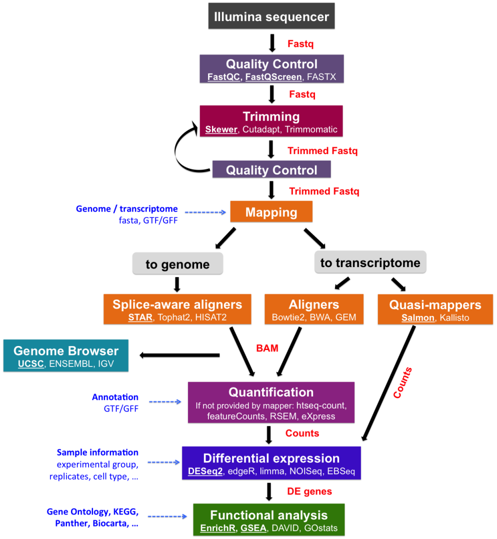
STAT588/BIOL588: Genomic Data Science
Lecture 17: Next Sequencing Data Anlaysis: Alignment (Chapter 5 in Gondro's book)

Dr. Yen-Yi Ho (hoyen@stat.sc.edu)

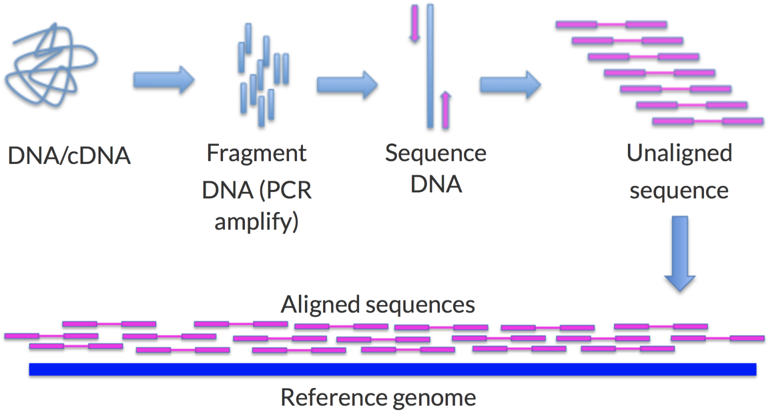
Next Generation Sequencing (NGS) Data Analysis: Alignment

- Alignment: BWT
- Simple DNA mapping
  - Bowtie2 and Samtools
  - R
- RNAseq
  - Workflow Example
  - RNAseq Normalization

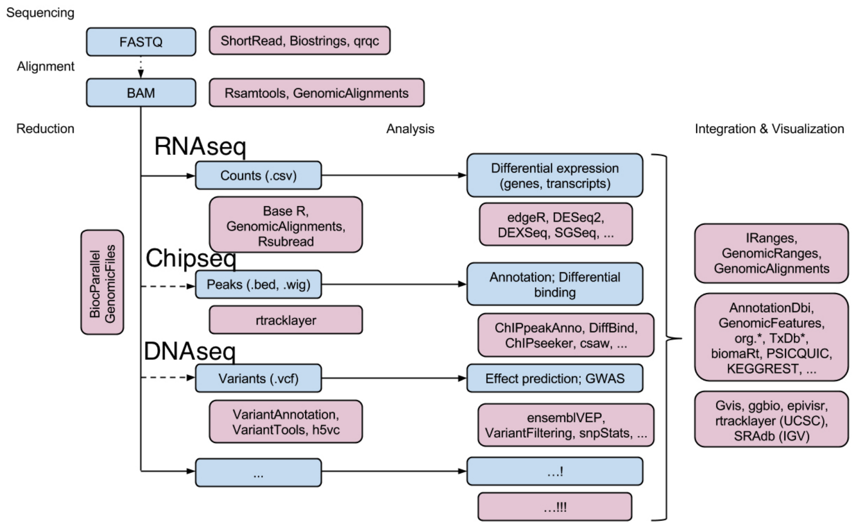
Work Flow


Illumina sequencer
Fastq
Quality Control
FastQC, FastQScreen, FASTX
Fastq
Trimming
Skewer, Cutadapt, Trimmomatic
Trimmed Fastq
Quality Control
Trimmed Fastq
Genome / transcriptome
fasta, GTF/GFF
Mapping
to genome
to transcriptome
Splice-aware aligners
STAR, Tophat2, HISAT2
Aligners
Bowtie2, BWA, GEM
Quasi-mappers
Salmon, Kallisto
Genome Browser
UCSC, ENSEMBL, IGV
BAM
Counts
Annotation
GTF/GFF
Quantification
If not provided by mapper: htseq-count, featureCounts, RSEM, eXpress
Counts
Sample information
experimental group, replicates, cell type, ...
Differential expression
DESeq2, edgeR, limma, NOISeq, EBSeq
DE genes
Gene Ontology, KEGG, Panther, Biocarta, ...
Functional analysis
EnrichR, GSEA, DAVID, GOstats

Mapping of Short Reads



Tools in R



Sequencing
FASTQ
ShortRead, Biostrings, qrqc
Alignment
BAM
Rsamtools, GenomicAlignments
Reduction
Analysis
Integration & Visualization
RNAseq
Counts (.csv)
Base R, GenomicAlignments, Rsubread
Differential expression (genes, transcripts)
edgeR, DESeq2, DEXSeq, SGSeq, ...
Chipseq
Peaks (.bed, .wig)
rtracklayer
Annotation; Differential binding
ChIPpeakAnno, DiffBind, ChIPseeker, csaw, ...
DNAseq
Variants (.vcf)
VariantAnnotation, VariantTools, h5vc
Effect prediction; GWAS
ensemblVEP, VariantFiltering, snpStats, ...
...
...!
...!!!
BiocParallel GenomicFiles
IRanges, GenomicRanges, GenomicAlignments
AnnotationDbi, GenomicFeatures, org.*, TxDb*, biomaRt, PSICQUIC, KEGGREST, ...
Gvis, ggbio, epivisr, rtracklayer (UCSC), SRAdb (IGV)

	Splice	Tool	Description
1	No	BWA	Burrows-Wheeler Transform algorithm based tool that accurately maps reads (up to 1 Mbp) to a given reference genome.
2	No	Bowtie2	Memory-efficient aligner for mapping very short reads (ranging from 50 to 100bp) to large genomes.
3	Yes	STAR	Spliced read aligner for de novo identification of novel splice junctions. STAR is significantly faster at read mapping compared with other sequence aligners.
4	Yes	Rsubread/ QuasR	Uses Rbowtie but can detect spliced junction

Table: Commonly used tools for short read alignment

“Next generation sequencing technology and genomewide data analysis: Perspectives for retinal research”, Progress in Retinal and Eye Research 55 (2016) 1e31.

Alignment

Bowtie is a program that aligns short reads to an indexed genome. Burrows Wheeler transform (BWT) and the FM index are implemented by Bowtie to perform read alignments.

- the FM index: Ferragina, P. and Manzini, G. (2000), "Opportunistic data structures with applications"
- BWT: Burrows, M. and Wheeler, D.J. (1994), "A Block-sorting lossless data compression algorithm".

Computing BWT: An Example

$$\text{GCCACC} \rightarrow \text{GCCACC\$}$$

Then get rotation matrix (M):

```
$GCCACC
C$GCCAC
CC$GCCA
ACC$GCC
CACC$GC
CCACC$G
GCCACC$
```

Computing BWT: An Example

Then sort the rotation matrix.

```
 F      L
 $GCCACC
 ACC$GCC
 C$GCCAC
↓CACC$GC
 CC$GCCA
 CCACC$G
 GCCACC$
```

We call the first column F and last column L. We can recover the whole sequence by storing only F and L. The index F and L can be stored in a very efficient manner (multiple Cs).

BWT: FM index

F	L
\$	C_0
A_0	C_1
C_0	C_2
C_1	C_3
C_2	A_0
C_3	G_0
G_0	\$

$$\overleftarrow{G_0\,C_3\,C_1\,A_0\,C_2\,C_0\,\$}$$

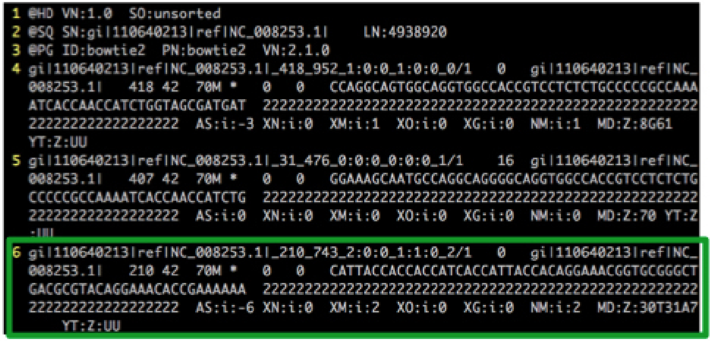
Alignment Algorithms

Before the alignment, it is important to know if the experiment was single-end or paired-end. The output of alignment step is commonly stored in SAM/BAM format.

- SAM file (.sam): Sequence Alignment/Map (SAM) is a text file that stores alignment information of reads to reference genome or given sequence. Some aligners such as STAR generate SAM file as an output of alignment process of short reads to reference genome. A SAM file includes a header section starting with @ character and alignment section consisting of multiple lines.
- BAM file (.bam): Binary Alignment/Map (BAM) is the binary version of SAM file. As SAM file does, BAM file stores alignment information of reads however BAM file is compressed (has smaller size) and more efficient in many sequencing analysis tools as it is compared to SAM file. A SAM file can be converted to a BAM file (or vice versa) with the help of SAMtools.

Headers

Alignments

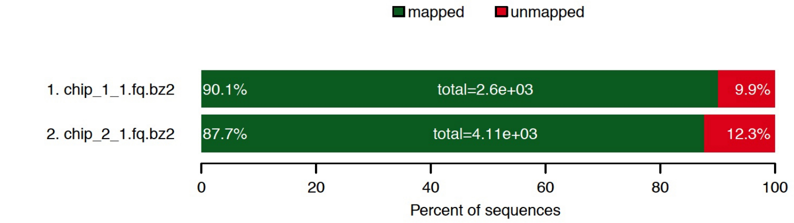


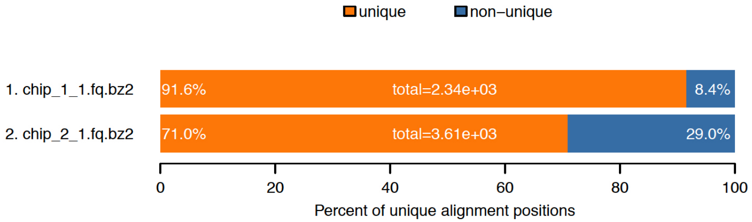
```
@HD VN:1.0  SO:unsorted
@SQ SN:gi|110640213|ref|NC_008253.1|    LN:4938920
@PG ID:bowtie2  PN:bowtie2  VN:2.1.0
gi|110640213|ref|NC_008253.1|_418_952_1:0:0_1:0:0_0/1   0   gi|110640213|ref|NC_
008253.1|   418 42  70M *   0   0   CCAGGCAGTGGCAGGTGGCCACCGTCCTCTCTGCCCCCGCCAAA
ATCACCAACCATCTGGTAGCGATGAT  2222222222222222222222222222222222222222222222222222
222222222222222222  AS:i:-3 XN:i:0  XM:i:1  XO:i:0  XG:i:0  NM:i:1  MD:Z:8G61
YT:Z:UU
gi|110640213|ref|NC_008253.1|_31_476_0:0:0_0:0:0_1/1    16  gi|110640213|ref|NC_
008253.1|   407 42  70M *   0   0   GGAAAGCAATGCCAGGCAGGGGCAGGTGGCCACCGTCCTCTCTG
CCCCCGCCAAAATCACCAACCATCTG  2222222222222222222222222222222222222222222222222222
222222222222222222  AS:i:0  XN:i:0  XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:70 YT:Z
:UU
gi|110640213|ref|NC_008253.1|_210_743_2:0:0_1:1:0_2/1   0   gi|110640213|ref|NC_
008253.1|   210 42  70M *   0   0   CATTACCACCACCATCACCATTACCACAGGAAACGGTGCGGGCT
GACGCGTACAGGAAACACCGAAAAAA  2222222222222222222222222222222222222222222222222222
222222222222222222  AS:i:-6 XN:i:0  XM:i:2  XO:i:0  XG:i:0  NM:i:2  MD:Z:30T31A7
    YT:Z:UU
```

Each row describes a single alignment of a raw read against the reference genome.
Each alignment has 11 mandatory fields, followed by any number of optional fields.

DNA Alignment in R

There are R packages (Rsubread, and QuasR) available for implementing alignment in R. These program uses Rbowtie. (R code examples in Lab17.R)





RNA-seq alignment

- Transcriptome: the collection of all transcripts that can be generated from a genome.
- $\approx$ 92-94% of human transcripts with more than one exon have alternatively spliced isoforms. (Wang, Sandberg, Luo et al. Nature 2008) $\rightarrow$ the same gene generates multiple mRNA transcripts.
- Due to splicing, there is a difference in aligning a set of reads from an RNA sample and a DNA sample.

RNA-seq alignment

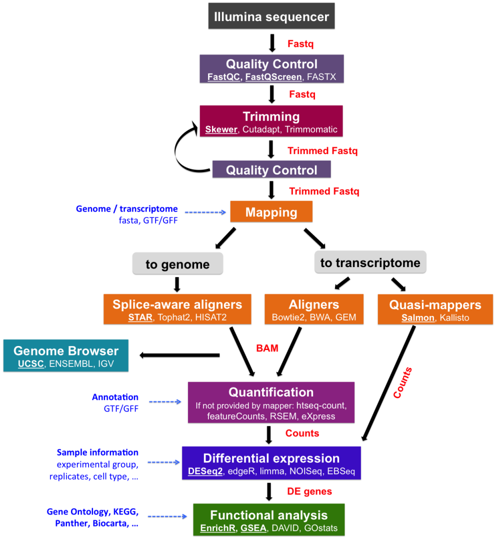
There are a number of algorithms that are designed to align a set of short reads to transcriptome.

- STAR
- TopHat
- MapSplice
- RSubread (R)
- QuasR (R)
- ...

R code using QuasR and RSubread are in Lab17.R. Linux command using STAR is also posted on the course website.

A list of software and comparison can be found at "Tools for mapping high-throughput sequencing data", Bioinformatics (2012) 28 (24): 3169-3177."

Work Flow



Illumina sequencer
Fastq
Quality Control
FastQC, FastQScreen, FASTX
Fastq
Trimming
Skewer, Cutadapt, Trimmomatic
Trimmed Fastq
Quality Control
Trimmed Fastq
Genome / transcriptome
fasta, GTF/GFF
Mapping
to genome
to transcriptome
Splice-aware aligners
STAR, Tophat2, HISAT2
Aligners
Bowtie2, BWA, GEM
Quasi-mappers
Salmon, Kallisto
Genome Browser
UCSC, ENSEMBL, IGV
BAM
Counts
Annotation
GTF/GFF
Quantification
If not provided by mapper: htseq-count, featureCounts, RSEM, eXpress
Counts
Sample information
experimental group, replicates, cell type, ...
Differential expression
DESeq2, edgeR, limma, NOISeq, EBSeq
DE genes
Gene Ontology, KEGG, Panther, Biocarta, ...
Functional analysis
EnrichR, GSEA, DAVID, GOstats

Getting Read Counts

There are several approaches to get read count from mapped RNA reads. The table below listed some of the popular approaches:

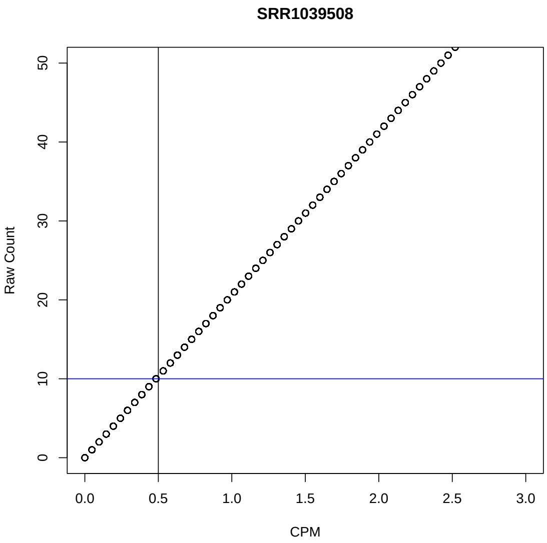
Function	Package	Framework	Output	DESeq2 input function
summarizeOverlaps	GenomicAligenments	R/Bioconductor	Summarized Experiment	DESeqDataSet
featureCounts	Rsubread	R/Bioconductor	matrix	DESeqDataSetFromMatrix
tximport	tximport	R/Bioconductor	list of matrices	DESeqDataSetFromTximport
htseq-count	HTSeq	Python	files	DESeqDataSetFromHTSeq

We use summarizeOverlaps and featureCounts as examples in Lab17.R

RNA-seq filtering

Gene with very low counts across all samples/libraries provide little evidence for differential expression. Instead of filtering genes based on raw counts directly, filtering could be based on count-per-million (CPM) due to differences in library sizes.

```
library("airway")
library("edgeR")
data(airway)
countMat<-assay(airway)
group<-airway$dex
#### CPM
myCPM<-cpm(countMat)
head(myCPM)
col1sum <- sum(countMat[,1])/1000000
countMat[1,1]/col1sum
thresh <- myCPM > 0.5
```



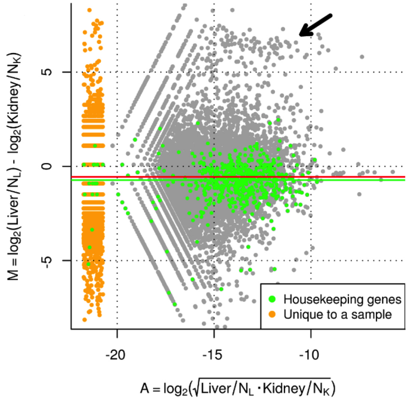
SRR1039508
Raw Count
CPM
0
10
20
30
40
50
0.0
0.5
1.0
1.5
2.0
2.5
3.0

RNAseq Normalzation

Factors that affect RNA-seq read counts.

- sequencing depth
- gene length
- composition

TMM (Trimmed mean of M values)



- One normalization factor per sample
- Compute M and A values for all genes.
- Discard gene with extreme M and A values. Then compute a weighted mean of M's for the rest of the genes.
- Assume most genes are not DE.

Example: Airway read counts

```
> library("airway")
> library("edgeR")
> data(airway)
> countMat<-assay(airway)
```

	1.bam	2.bam	3.bam	4.bam	5.bam	...
ENSG00000009724	38	28	66	24	42	
ENSG00000116649	1004	1255	1122	1313	1100	
ENSG00000120942	218	256	233	252	269	
...						

Normalization in edgeR

```
> countMat<-assay(airway)
> group<-airway$dex
> group<-relevel(group, "untrt")
> y<-DGEList(counts=countMat, group=group)
> y<-calcNormFactors(y)
> y
$samples
           group lib.size norm.factors
SRR1039508 untrt 20637971    1.0553567
SRR1039509   trt 18809481    1.0291947
SRR1039512 untrt 25348649    0.9832690
SRR1039513   trt 15163415    0.9490081
SRR1039516 untrt 24448408    1.0255871
SRR1039517   trt 30818215    0.9729022
SRR1039520 untrt 19126151    1.0308785
SRR1039521   trt 21164133    0.9592055
```